Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by dev on Tue, 12 Sep 2006 14:03:27 GMT
View Forum Message <> Reply to Message

Herbert Poetzl wrote:

>>><<<< such checks for CAP_SYS_ADMIN mean that we can't use
>>>copy_xxx/clone_xxx functions directly
>>><<<< from OpenVZ code, since VE creation is done with dropped
>>>capabilities already.
>
>
> is there a good reason for doing so?
> I mean, Linux-VServer for example drops the capabilities
> at the end of initialization, right before spawning the
> guest init (or running the guest's runlevel scripts)
yes, there is a security reason.
default set of capabilities is saved on VE creation to
ve->cap_default. This is used to make sure that on VE 'enter'
a process moved between contexts won't leak capabilities to VE.

So when VE is created it should be known already which caps
to use.

>>><<<< (user level tools decide which capabilities should be granted
>>>to VE, so CAP_SYS_ADMIN
>>><<<< is not normally granted :))
>>><<<< Can we move capability checks into some more logical place
>>>which deals with user, e.g. sys_unshare()?

Kirill

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Tue, 12 Sep 2006 14:24:22 GMT
View Forum Message <> Reply to Message

On Tue, Sep 12, 2006 at 06:07:20PM +0400, Kirill Korotaev wrote:
> Herbert Poetzl wrote:
>
> >>><<<< such checks for CAP_SYS_ADMIN mean that we can't use
> >>>copy_xxx/clone_xxx functions directly
> >>><<<< from OpenVZ code, since VE creation is done with dropped
> >>>capabilities already.
> > >
> > >
> > > is there a good reason for doing so?
> > > I mean, Linux-VServer for example drops the capabilities

> > at the end of initialization, right before spawning the
> > guest init (or running the guest's runlevel scripts)

> yes, there is a security reason.
> default set of capabilities is saved on VE creation to
> ve->cap_default. This is used to make sure that on VE 'enter'
> a process moved between contexts won't leak capabilities to VE.

well, we (Linux-VServer) can probably help you here:

we figured some time ago, that applying the capability
restriction to the capability set has two disadvantages
when done at guest startup

 a) there is a small chance that a process could
    unintentionally get a higher capability from
    outside (host system) after startup

 b) changes to the capability set will only affect
    newly created processes, which typically requires
    a guest or service restart

we therefore decided to have a capability 'mask' for
each guest, which is applied to the current/actual
capabilities whenever the caps are checked, and of
course, this mask can be set at creation time too,
as it does not affect the creating process until
the setup has finished

HTH,
Herbert

> So when VE is created it should be known already which caps
> to use.
>
> >>><<<< (user level tools decide which capabilities should be granted
> >>>to VE, so CAP_SYS_ADMIN
> >>><<<< is not normally granted :))
> >>><<<< Can we move capability checks into some more logical place
> >>>which deals with user, e.g. sys_unshare()?
>
> Kirill
> _____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers